INTER-SYSTEM EXCLUSIVE CONTROL SYSTEM

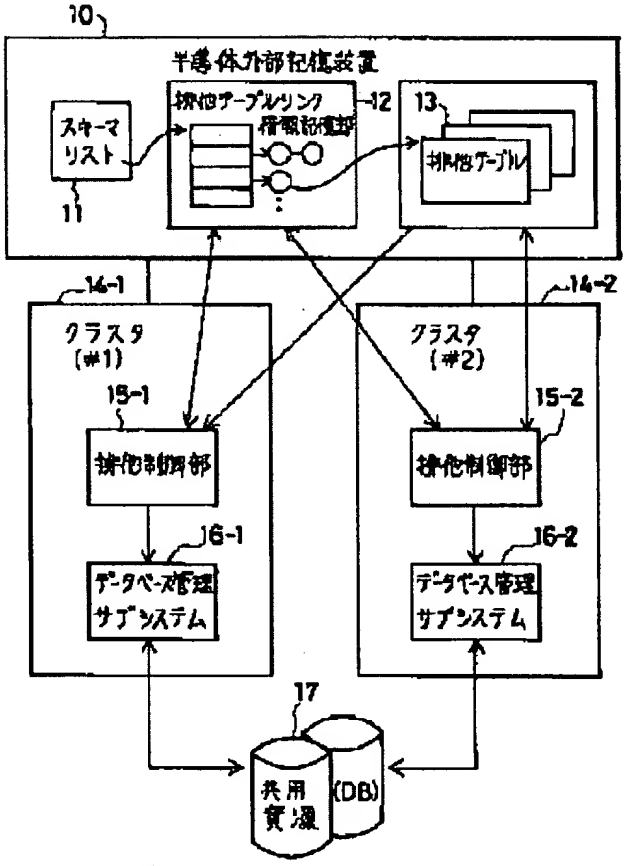
Patent number: JP5128072
Publication date: 1993-05-25
Inventor: SHIGA KOICHI; YANASE YUKIYOSHI
Applicant: FUJITSU LTD
Classification:
- international: G06F15/16
- european:
Application number: JP19910289031 19911106
Priority number(s): JP19910289031 19911106

Also published as:
US5559979 (A1)

Report a data error here

Abstract of **JP5128072**

PURPOSE:To execute efficient and rapid exclusive control in an inter-system exclusive control system for executing the exclusive control of resources shared by plural clusters in a composite system. CONSTITUTION:Necessary exclusive management information corresponding to respective resources is stored in respective exclusive tables 13 stored in a semiconductor external storage device 10. Address information corresponding to respective tables 13 is stored in an exclusive table link information storing part 12 together with resource identification (ID) information and these information is integrated to capacity less than the whole capacity of the tables 13. At the time of inputting an exclusive request, exclusive control parts 15-1, 15-2 read out information from the storing part 12, judge the existence of another exclusive request, and when another exclusive request exists, access to the table 13 concerned in accordance with the address information of the table 13 and execute exclusive processing corresponding to the contents of the table 13.



(19)日本国特許庁（ＪＰ）　(12)公開特許公報(A)　(11)特許出願公開番号

特開平5－128072

(43)公開日　平成5年(1993)5月25日

(51)Int.Cl.[5]	識別記号	庁内整理番号	Ｆ Ｉ	技術表示箇所
Ｇ０６Ｆ　15/16	３４０　Ａ	8840－5Ｌ		

審査請求　未請求　請求項の数１(全 11 頁)

(21)出願番号　特願平3－289031

(22)出願日　平成３年(1991)11月６日

(71)出願人　000005223
富士通株式会社
神奈川県川崎市中原区上小田中1015番地
(72)発明者　志賀　浩一
神奈川県川崎市中原区上小田中1015番地
富士通株式会社内
(72)発明者　柳瀬　幸好
神奈川県川崎市中原区上小田中1015番地
富士通株式会社内
(74)代理人　弁理士　小笠原　吉義　(外２名)

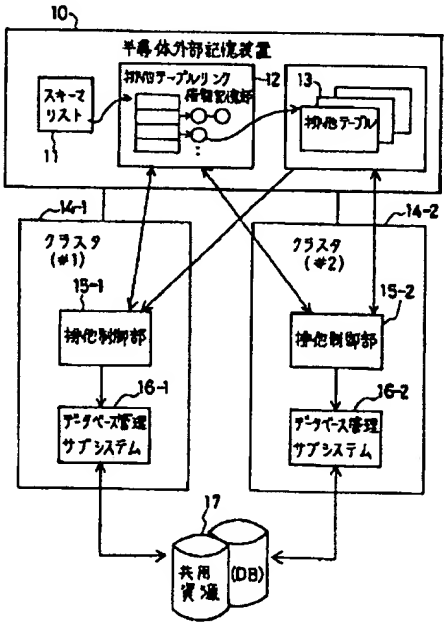
(54)【発明の名称】　システム間排他制御方式

(57)【要約】

【目的】本発明は複合システムにおいてクラスタ間14-1,14-2 で共用する資源17の排他制御を行うシステム間排他制御方式に関し，効率的で高速な排他制御を実現する手段を提供することを目的とする。

【構成】半導体外部記憶装置10内の排他テーブル13には，各資源対応に必要な排他管理情報を記憶する。排他テーブルリンク情報記憶部12には，排他テーブル13に対するアドレス情報を資源の識別情報と共に保持させ，これを排他テーブル13の全体よりも大きくない容量にまとめる。排他制御部15-1,15-2 は，排他要求に対して，排他テーブルリンク情報記憶部12の情報を読み込み，他の排他要求の有無を判定すると共に，他の排他要求がある場合に，その排他テーブル13に対するアドレス情報により該当する排他テーブル13をアクセスし，排他テーブルの内容に応じた排他処理を行うように構成される。

本発明の原理説明図



1

【特許請求の範囲】

【請求項1】　通信可能に結合された複数の計算機システム(14)と，各計算機システム間で排他制御の対象となる共用資源(17)と，各計算機システムから共用される半導体外部記憶装置(10)とを備えた複合システムにおけるシステム間排他制御方式において，前記半導体外部記憶装置内に，システム間排他制御の対象となっている各資源対応に，少なくとも使用元情報およびその資源に対するアクセスを逐次化するための排他情報を含む排他管理情報を記憶する排他テーブル(13)と，前記排他テーブルに対するアドレス情報を資源の識別情報と共に保持するテーブルであって，前記排他テーブルの全体よりも大きくない容量にまとめられた排他テーブルリンク情報記憶部(12)とを有し，各計算機システムごとに，共用資源に対する排他要求に対して，前記半導体外部記憶装置から前記排他テーブルリンク情報記憶部の情報を読み込み，当該共用資源に対する他の排他要求の有無を判定すると共に，他の排他要求がある場合に，その排他テーブルに対するアドレス情報により該当する排他テーブルをアクセスし，排他テーブルの内容に応じた排他処理を行う排他制御部(15)を備えたことを特徴とするシステム間排他制御方式。

【発明の詳細な説明】

【0001】

【産業上の利用分野】本発明は，複合システムにおいて，計算機システム間で共用する資源の排他制御を行うシステム間排他制御方式に関する。

【0002】複数の計算機システムを通信路で結合した形態のシステムを，複合システムと呼ぶ。複合システムを構成する各計算機システムを，以後，クラスタと呼ぶ。計算機システムの多様化，大規模化により，最近，このような複合システムが用いられるようになってきたが，クラスタ間で共用するデータベースなどの共用資源に対するアクセスを，クラスタ間で排他制御する必要があり，そのオーバヘッドをできるだけ少なくする技術が要求されている。

【0003】

【従来の技術】図6は従来技術の説明図である。従来，データベースのクラスタ間排他制御方式としては，図6の（イ）に示すようなハードウェアのリザーブ・リリーズを利用した方式や，図6の（ロ）に示すようなクラスタ間通信を利用した方式が用いられている。

【0004】ハードウェアのリザーブ・リリーズを利用する方式では，図6の（イ）に示すように，クラスタ14－1がデータベース63にアクセスするとき，まず排他機構60－1によって，データベース63が配置されているボリュームに対して，リザーブの入出力命令を発行し（図示①），そのリザーブが成功したならば，データベース管理サブシステム16－1によって，データベース63にアクセス（図示②）する。

2

【0005】同時にクラスタ14－2でデータベース63に対するアクセス要求があっても，データベース63のボリュームが，クラスタ14－1にリザーブされている間は，排他機構60－2によるリザーブは成功しないので，クラスタ14－2からのデータベース63に対するアクセスは禁止される。

【0006】また，クラスタ間通信を利用する方式では，図6の（ロ）に示すように，クラスタ14－1の応用処理部61－1からアクセス要求（図示①）があると，排他機構60－1は，通信機構62－1，62－2を用いて排他要求（図示②）を他のクラスタ14－2に伝える。クラスタ14－2の排他機構60－2は，自クラスタがデータベース63を現在使用しているかどうかにより，排他可否の判定処理（図示③）を行い，その排他結果（図示④）を通信機構62－2，62－1を経由して排他機構60－1に通知する。排他機構60－1は，排他成功であれば，結果をデータベース管理サブシステム16－1に通知し，データベース管理サブシステム16－1はデータベース63にアクセス（図示⑤）する。

【0007】

【発明が解決しようとする課題】図6の（イ）に示すようなハードウェアのリザーブ・リリーズを利用する方式では，各アクセス要求ごとにリザーブ・リリーズが必要となるため，排他制御のオーバヘッドが大きく，またボリューム単位の排他となるため，排他の単位が大きくなり，クラスタ間の排他待ちが発生しやすいという問題がある。

【0008】また，図6の（ロ）に示すようなクラスタ間通信を利用する方式では，複合システムのクラスタ数の増大に比例して通信のオーバヘッドが増加するという問題，および通信路の性能以上にトランザクション処理性能を向上できないという問題がある。

【0009】本発明は上記問題点の解決を図り，排他の単位をデータベースの格納構造に対応した範囲で任意に設定できるようにし，かつ排他制御のオーバヘッドを少なくして，効率的で高速な排他制御を実現する手段を提供することを目的としている。

【0010】

【課題を解決するための手段】図1は本発明の原理説明図である。図1において，10は半導体メモリを利用した高速アクセスが可能な半導体外部記憶装置，11はデータベースの構造定義体であるスキーマの情報を持つスキーマリスト，12は排他テーブルリンク情報記憶部，13は排他テーブル，14－1，14－2は各々CPUおよび主記憶を持つ計算機システムからなるクラスタ，15－1，15－2はデータベースなどの資源に対する排他要求を処理する排他制御部，16－1，16－2はデータベースに対するアクセス機能を提供するデータベース管理サブシステム，17はクラスタ14－1，14

3

－2が共用するデータベースなどの共用資源を表す。

【0011】半導体外部記憶装置10は，各クラスタ14－1，14－2からアクセス可能となっており，クラスタ14－1，14－2は複合システムを構成している。本発明では，この半導体外部記憶装置10内に，クラスタ間排他制御の対象となっている各資源対応に，少なくとも使用元情報およびその資源に対するアクセスを逐次化するための排他情報を含む排他管理情報を記憶する排他テーブル13を持つ。

【0012】また，半導体外部記憶装置10内に，排他テーブル13に対するアドレス情報を資源の識別情報と共に保持するテーブルであって，排他テーブル13の全体よりも大きくない容量にまとめられた排他テーブルリンク情報記憶部12を持つ。

【0013】排他制御部15－1（排他制御部15－2も同様）は，共用資源17に対する排他要求があると，半導体外部記憶装置10から排他テーブルリンク情報記憶部12の内容をクラスタ14－1の主記憶に読み込み，その共用資源17に対する他の排他要求の有無を，読み込んだ情報により判定すると共に，他の排他要求がある場合には，その排他テーブル13に対するアドレス情報により該当する排他テーブル13をアクセスし，排他テーブル13の内容に応じた排他処理を行う。

【0014】例えば共用資源17がデータベースである場合，排他テーブルリンク情報記憶部12はスキーマリスト11からポイントされる。排他テーブル13の有無は，排他テーブルリンク情報記憶部12によって判るので，すべての排他テーブル13を検索することなく，半導体外部記憶装置10に対する少ないアクセス回数で排他処理を実行することができる。

【0015】排他制御部15－1は，排他に成功すると，必要な情報を排他テーブル13に設定し，それを排他テーブルリンク情報記憶部12からポイントして，データベース管理サブシステム16－1に排他成功を通知する。データベース管理サブシステム16－1は，要求に応じた共用資源17に対するアクセスを行う。

【0016】

【作用】本発明では，クラスタ間排他制御表を，排他テーブルリンク情報記憶部12と排他テーブル13に分け，高速アクセスが可能な半導体外部記憶装置10に配置する。これをクラスタ間で排他的に参照・更新することにより，排他制御を実現する。

【0017】特に，半導体外部記憶装置10に対するアクセスを少なくするため，読み込み処理および書き込み処理の逐次化に必要な情報は，排他テーブル13に格納し，排他テーブルリンク情報記憶部12には，排他テーブル13へのポインタ情報を格納する。そして，排他テーブルリンク情報記憶部12を排他テーブル13の全体よりも大きくならないようにまとめる。したがって，例えば1回のアクセスで排他テーブルリンク情報記憶部12を読み込むだけで，該当する排他テーブル13の有無，すなわち占有済みトランザクションの有無などがわかると共に，該当する排他テーブル13を直ちにアクセスできるようになる。

4

【0018】

【実施例】以下，複合システムで共用するデータベースに関する排他制御について，本発明の実施例を説明する。

【0019】図2および図3は本発明の実施例で用いるテーブル構成図，図4は本発明の実施例に係るテーブル関連図を示す。図1に示すスキーマリスト（SCLT）11は，例えば図2の（イ）に示す構造になっている。スキーマリスト11は，ヘッダー部と各スキーマ対応のエントリ部とから構成され，ヘッダー部には，リストの全体の長さ，エントリ数，エントリ長が設定される。1つのエントリ部は，スキーマ名と，一般データセットであるとかリレーショナル・データベースであるとかいった資源種別名と，当該スキーマに関する排他テーブルハッシュ（STHS）のアドレスを管理する。

【0020】図1に示す排他テーブルリンク情報記憶部12は，図2の（ロ）に示す排他テーブルハッシュ20と，図3の（イ）に示すリンクテーブル（LNKT）30とから構成される。排他テーブルハッシュ20は，ヘッダー部と複数のエントリ部とからなり，ヘッダー部は，このテーブルの全体長とエントリ数とエントリ長の情報を持つ。1つのエントリ部は，例えばサブスキーマなどの排他制御の単位となる資源を識別する資源識別子と，リンクテーブル30へのアドレス情報を持つ。

【0021】リンクテーブル30は，図3の（イ）に示すように，ヘッダー部と複数のエントリ部から構成され，ヘッダー部は，リンクテーブルの全体長と未使用の先頭のリンクテーブルエントリへのアドレス情報を持つ。1つのエントリ部は，クラスタ間排他テーブル部31のエントリ部へのアドレスと，次のリンクテーブルへのアドレスの情報を持つ。この次リンクテーブルアドレスによって，各エントリ部はキューイング可能になっている。

【0022】クラスタ間排他テーブル部31は，図3の（ロ）に示すように，ヘッダー部と複数のエントリ部とから構成され，ヘッダー部は，このテーブルの全体長と未使用の先頭の排他テーブルエントリへのアドレス情報を持つ。1つのエントリ部は，図1に示す排他テーブル13に相当し，1つのトランザクションを管理する。詳しくは，トランザクションを識別するトランザクション識別子，いわゆるS（共用）モード，X（占有）モードというような各種の排他レベル情報および資源識別子を管理する。

【0023】上記各テーブルの関連は，図4に示すようになっている。排他制御に関するテーブルは，図4に示す排他テーブルリンク情報記憶部12のハッシュ部と，

5

クラスタ間排他テーブル部３１のデータ部とからなる。このデータ部には，排他テーブル１３－１，１３－２，…が含まれる。各排他テーブル１３－１，…はリンクテーブル３０－１，…を介して，排他テーブルハッシュ２０のエントリに対応づけられる。

【００２４】例えば，ある共用資源（ＢＬＯＣＫ２）が既にトランザクション（ＴＲＮ１）によって占有されている場合には，排他テーブルハッシュ２０の該当エントリから，リンクテーブル３０－３へのポインタが張られており，このリンクテーブル３０－３から排他テーブル１３－１がポイントされる。排他テーブル１３－１には，占有要求を発生させたＴＲＮ１のトランザクション識別子，排他レベル等が管理されているので，これを参照してクラスタ間のトランザクション排他処理が可能となる。

【００２５】排他テーブルリンク情報記憶部１２を排他テーブルハッシュ２０とリンクテーブル３０－１，…に分け，各リンクテーブルをキューで管理する構成とすることにより，排他テーブルハッシュ２０の１つのエントリを複数の排他テーブル１３に対応づけるときにも，少ないメモリ量で済むようになっている。

【００２６】１つのスキーマに対応する排他テーブルハッシュ２０とリンクテーブル３０－１，…を合わせたハッシュ部の大きさは，半導体外部記憶装置１０に対する１回のアクセスで読み込み可能な大きさとする。また，１つのヘッダー部とエントリ部を合わせたクラスタ間排他テーブル部３１の大きさも，半導体外部記憶装置１０のアクセス単位に合わせる。これにより，半導体外部記憶装置１０に対するアクセス回数を最小にすることができる。

【００２７】図５は，本発明の実施例による排他制御部の処理フローを示す。以下，図５に示す処理(a) ～(r) に従って，排他要求に対する排他制御部の処理を説明する。

(a) 応用プログラムから，クラスタ間共用データベースのレコードに対するアクセス（参照／更新）命令が発行されると，データベース管理サブシステムは排他制御部に対して排他処理を依頼する。排他制御部は，まず排他処理を依頼された資源が，クラスタ間の共用資源であるかどうかを判定する。

【００２８】(b) 排他対象の資源がクラスタ間の共用資源でない場合，クラスタ内のトランザクション排他処理を行う。この処理については従来と同様でよいので，ここでの詳しい説明は省略する。

【００２９】(c) クラスタ間の共用資源である場合，データベース管理サブシステムから通知されたスキーマ名と資源種別名をキーにして，スキーマリストを検索し，該当する排他テーブルハッシュのアドレスを求める。なお，この例ではスキーマリストは半導体外部記憶装置にあるが，あらかじめ主記憶に読み込んでおくことにより，高速な検索を可能としている。

6

【００３０】(d) 半導体外部記憶装置における排他制御のテーブル（排他テーブルリンク情報記憶部および排他テーブル）に対して，参照・更新の競合を避けるために，クラスタ間排他ロックを取得する。このクラスタ間排他ロックは，例えば周知のコンペア・アンド・スワップ命令のような命令によって取得することができる。

【００３１】(e) クラスタ間排他ロックを取得したならば，排他制御表のハッシュ部，すなわち排他テーブルリンク情報記憶部を主記憶に読み込む。

(f) そして，資源識別子等をキーとする所定のハッシュ関数の適用または排他テーブルハッシュの検索により，排他対象の資源に関する排他テーブルハッシュエントリのアドレスを求める。

【００３２】(g) 排他テーブルハッシュエントリが求まったならば，その中のリンクテーブルアドレスが０かどうかを調べる。０の場合，この資源に対して排他処理が必要な他の排他要求は発行されていないので，処理(l) へ進む。

【００３３】(h) リンクテーブルアドレスが０でない場合，そのアドレスによってリンクテーブルをアクセスし，リンクテーブルからポイントされる排他テーブルを，半導体外部記憶装置から主記憶に読み込む。

【００３４】(i) 読み込んだ排他テーブルに設定されている排他レベルと，今回の要求に係る排他レベルとが，排他的かどうかを調べる。この排他レベルによる排他チェックについては，従来から種々のものが知られているので，ここでの詳しい説明は省略する。排他的でない場合，処理(k) へ進む。

【００３５】(j) 排他チェックにより，排他的であることがわかった場合，今回の排他要求を却下し，要求元へ不成功を通知する。または前の排他が解除されるまで，今回の排他要求を待たせる待ち制御を行う。なお，待ち制御には，例えばリンクテーブルを利用することができる。

【００３６】(k) 排他テーブルに設定されている排他レベルが排他的でない場合，現在のリンクテーブルにおける次のリンクテーブルアドレスを調べ，処理(g) 以下の処理を同様に繰り返す。

【００３７】(l) クラスタ間排他テーブル部のヘッダー部を参照し，未使用の排他テーブルのエントリを切り出す。

(m) クラスタ識別子，トランザクション識別子，排他レベル，資源識別子などの排他管理情報を，排他テーブルに設定する。

【００３８】(n) リンクテーブルのヘッダー部を参照し，未使用のリンクテーブルエントリを切り出す。

(o) 切り出したリンクテーブルのエントリから，処理(m) で設定した排他テーブルをポイントする。

【００３９】(p) そのリンクテーブルを，該当する排他

7

テーブルハッシュのエントリからキューイングされる最終のリンクテーブルに，キューイングする。

(q) 処理(l) ないし処理(p) を主記憶上で行った場合には，その更新内容を半導体外部記憶装置に書き戻し，半導体外部記憶装置の排他制御表に対する排他ロックを解放する。

【0040】(r) データベース管理サブシステムに排他処理の正常終了を通知し，排他処理を終了する。

【0041】

【発明の効果】以上説明したように，本発明によれば，従来のハードウェアのリザーブ・リリーズによるクラスタ間排他制御方式やクラスタ間通信による排他制御方式に比較して，クラスタ間排他処理を，複合システムを構成するクラスタ数に影響されないオーバヘッドで行うことができるようになる。

【0042】特に，ボリューム単位というような大きな排他の単位ではなく，例えばデータベースの格納構造に対応した範囲でクラスタ間排他を実現することができ，排他待ちの時間を短縮することが可能である。また，半導体外部記憶装置に配置する排他制御のための制御表を，排他テーブルリンク情報記憶部と実際の排他管理情報が設定される排他テーブルとに分けているため，半導体外部記憶装置に対するアクセス回数を少なくすることができ，排他処理の高速化が可能になる。

8

【0043】したがって，特定のデータにアクセスが集中するような業務に対しても，極端にレスポンスを悪化させることなく，負荷分散によって複合システムを有効に活用することができるようになる。

【図面の簡単な説明】

【図1】本発明の原理説明図である。

【図2】本発明の実施例で用いるテーブル構成図である。

【図3】本発明の実施例で用いるテーブル関連図である。

【図4】本発明の実施例に係るテーブル関連図である。

【図5】本発明の実施例による排他制御部の処理フローを示す図である。

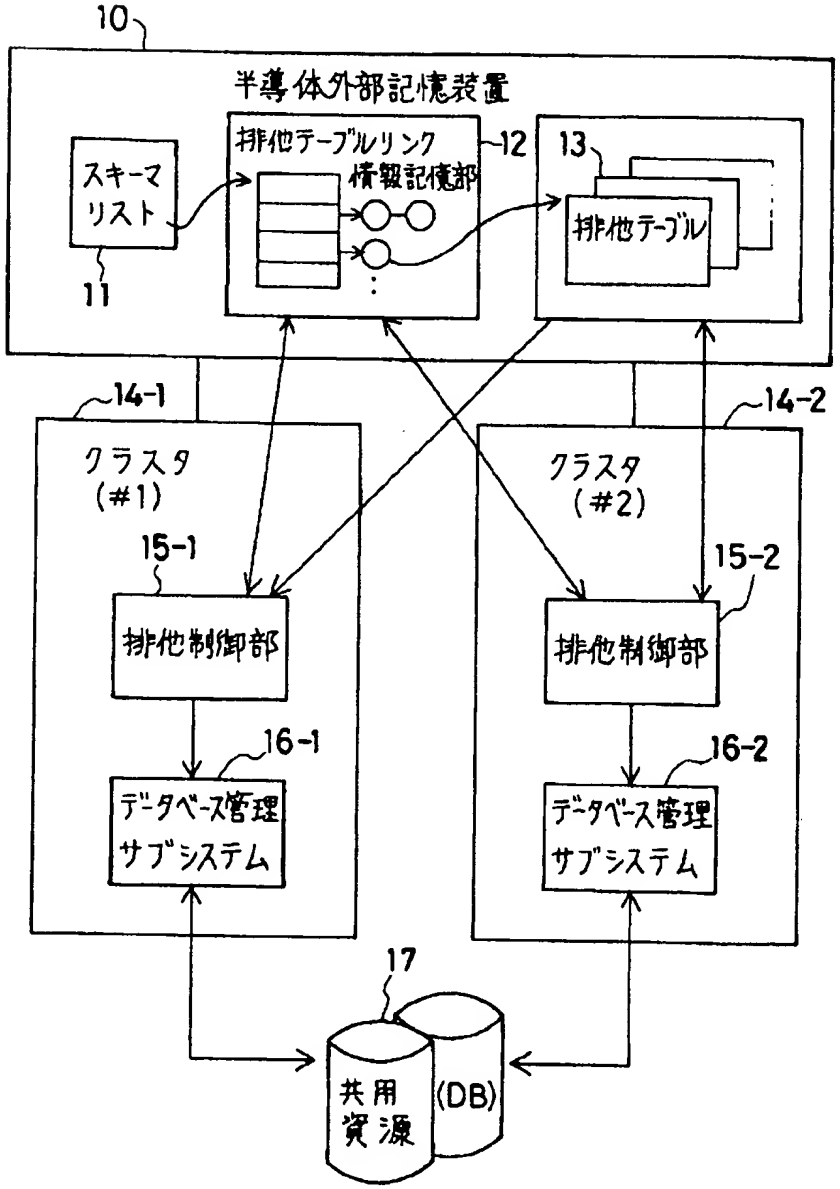
【図6】従来技術の説明図である。

【符号の説明】

10　半導体外部記憶装置
11　スキーマリスト
12　排他テーブルリンク情報記憶部
13　排他テーブル
14－1，14－2　クラスタ
15－1，15－2　排他制御部
16－1，16－2　データベース管理サブシステム
17　共用資源

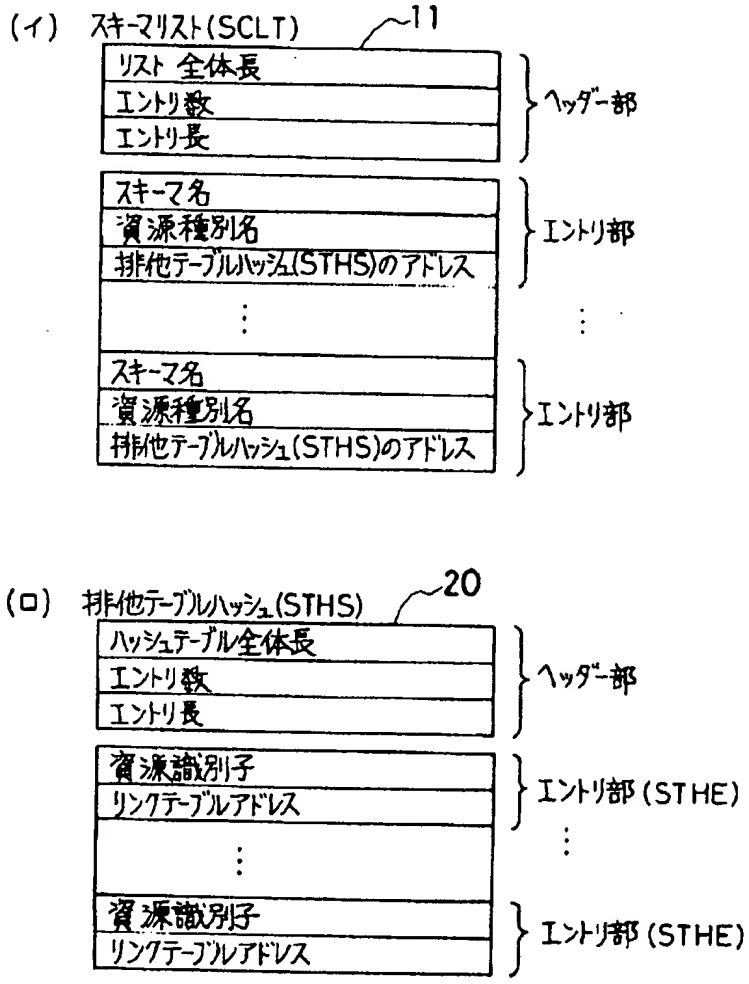
【図1】

本発明の原理説明図



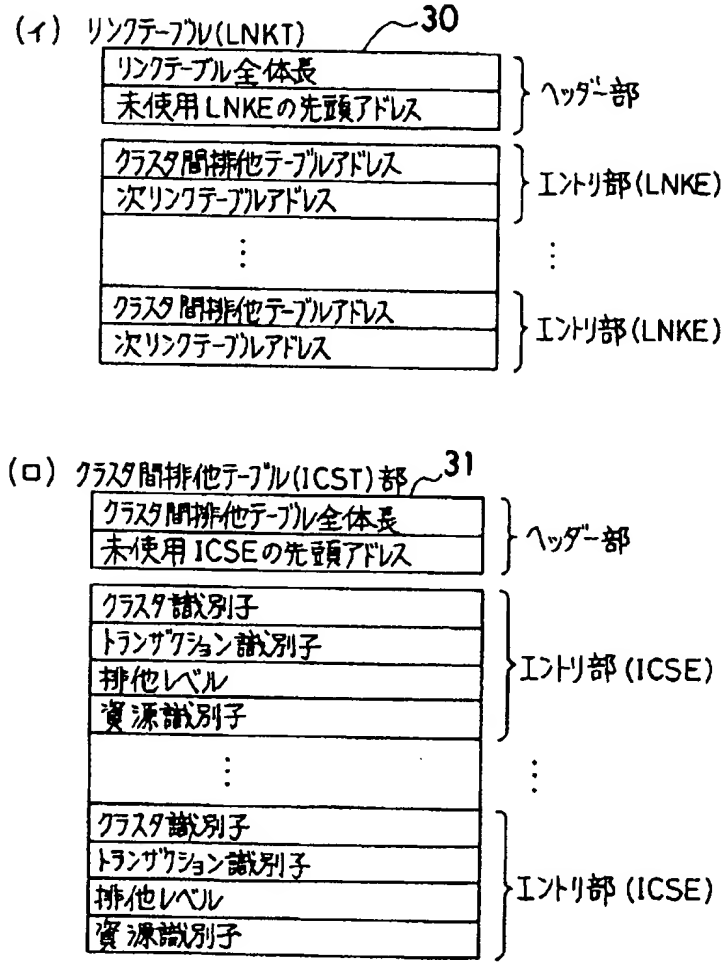
【図2】

テーブル構成図


(イ) スキーマリスト(SCLT)
11
リスト 全体長
エントリ数
エントリ長
ヘッダー部
スキーマ名
資源種別名
排他テーブルハッシュ(STHS)のアドレス
エントリ部
スキーマ名
資源種別名
排他テーブルハッシュ(STHS)のアドレス
エントリ部
(ロ) 排他テーブルハッシュ(STHS)
20
ハッシュテーブル全体長
エントリ数
エントリ長
ヘッダー部
資源識別子
リンクテーブルアドレス
エントリ部(STHE)
資源識別子
リンクテーブルアドレス
エントリ部(STHE)

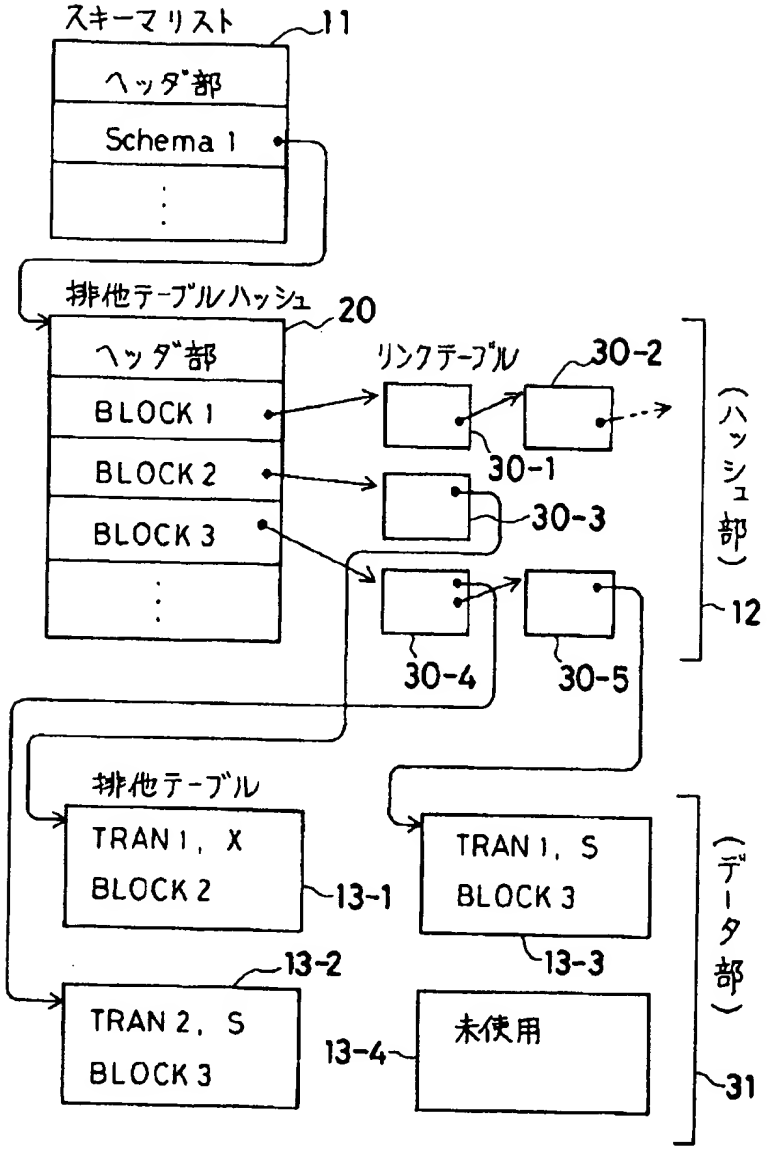
【図3】

テーブル関連図


(イ) リンクテーブル(LNKT)
30
リンクテーブル全体長
未使用LNKEの先頭アドレス
ヘッダー部
クラスタ間排他テーブルアドレス
次リンクテーブルアドレス
エントリ部(LNKE)
クラスタ間排他テーブルアドレス
次リンクテーブルアドレス
エントリ部(LNKE)
(ロ) クラスタ間排他テーブル(ICST)部
31
クラスタ間排他テーブル全体長
未使用ICSEの先頭アドレス
ヘッダー部
クラスタ識別子
トランザクション識別子
排他レベル
資源識別子
エントリ部(ICSE)
クラスタ識別子
トランザクション識別子
排他レベル
資源識別子
エントリ部(ICSE)

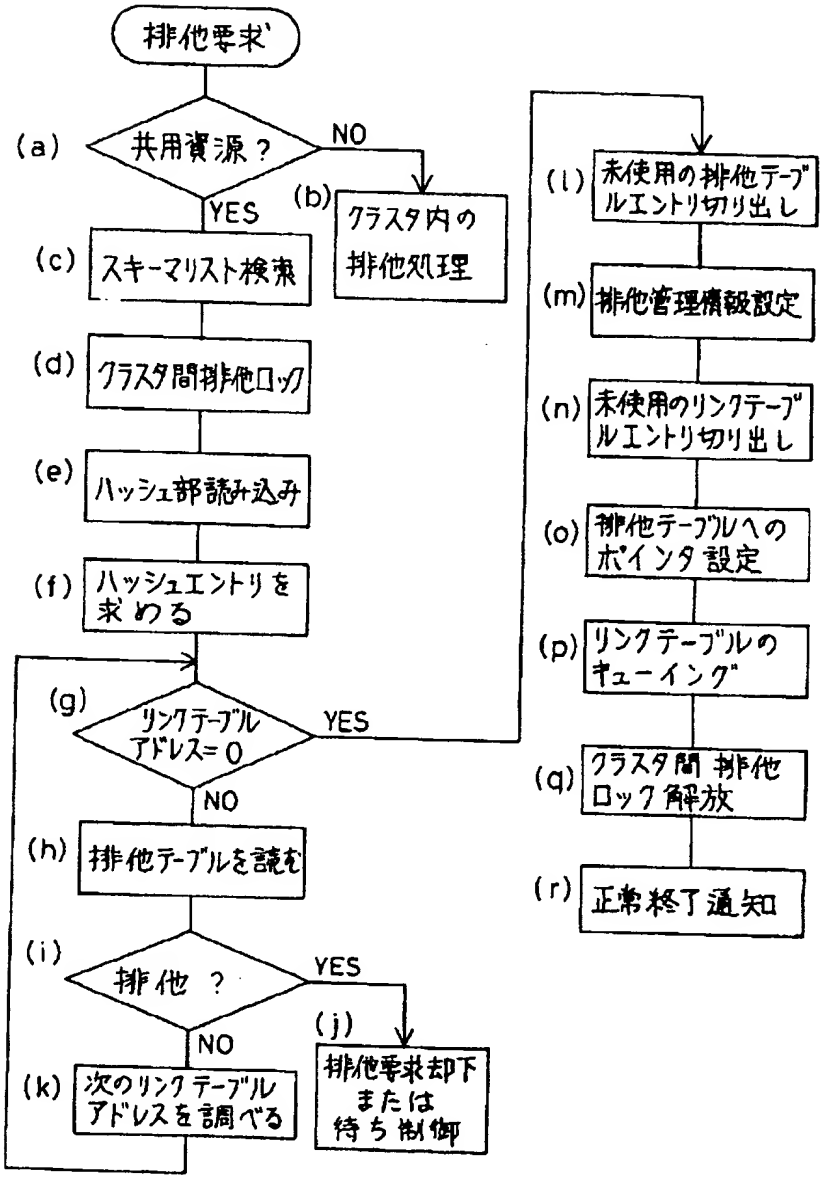
【図4】

テーブル関連図


スキーマリスト
11
ヘッダ部
Schema 1
排他テーブルハッシュ
20
ヘッダ部
BLOCK 1
BLOCK 2
BLOCK 3
リンクテーブル
30-1
30-2
30-3
30-4
30-5
(ハッシュ部)
12
排他テーブル
TRAN 1, X
BLOCK 2
13-1
TRAN 1, S
BLOCK 3
13-3
13-2
TRAN 2, S
BLOCK 3
13-4
未使用
(データ部)
31

【図5】

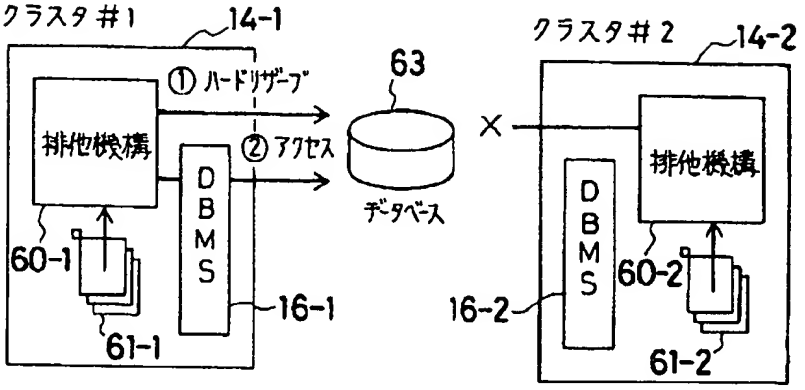
排他制御部の処理フロー



【図6】

従来技術の説明図

(イ)



(ロ)